Imputing Phenotypes for Genome-wide Association Studies

Imputing phenotypes for genome-wide association studies

Farhad Hormozdiari[1], Eun Yong Kang[1], Michael Bilow[1], Eyal Ben David[2]
Chris Vulpe[3], Stela McLachlan[4], Aldons J. Lusis[5,6], Buhm Han[7], Eleazar Eskin [1,6,‡‡]

1 Department of Computer Science, University of California, Los Angeles

2 Department of Genetics, Hebrew University of Jerusalem

3 Department of Nutritional Science and Toxicology, University of California Berkeley

4 Centre for Population Health Sciences, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh

5 Department of Medicine, University of California Los Angeles

6 Department of Human Genetics, University of California, Los Angeles

7 Department of Convergence Medicine, University of Ulsan College of Medicine & Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

‡‡ corresponding author : eeskin@cs.ucla.edu

Abstract

Genome-wide association studies (GWAS) have been successful in detecting variants correlated with phenotypes of clinical interest. However, the power to detect these variants depends on the number of individuals whose phenotypes are collected. Thus, for phenotypes that are difficult to collect, the sample size may be insufficient to achieve the desired statistical power. Often, while the phenotype of interest is difficult to collect, surrogate phenotypes or related phenotypes are easier to collect and have already been collected in very large samples. In this paper, we take advantage of these additional related phenotypes to impute the phenotype of interest or target phenotype and then perform association analysis. Our approach leverages the correlation structure between phenotypes to perform the imputation. The correlation structure can be estimated from a smaller complete dataset for which both the target and related phenotypes have been collected. Under some assumptions the statistical power can be computed analytically given the correlation structure of the phenotypes used in imputation. In addition,

our method can impute the summary statistic of the target phenotype as a weighted linear combination of the summary statistics of related phenotypes. Thus, our method is applicable to datasets for which we only have access to summary statistics and not the raw genotypes. We illustrate our approach by analyzing associated loci to triglycerides (TG), body mass index (BMI), and systolic blood pressure (SBP) in the Northern Finland Birth Cohort dataset.

1 Introduction

In genome-wide association studies (GWAS), investigators collect genotypes and phenotypes from a set of individuals and then perform a series of statistical tests to identify variants that are significantly associated with the phenotype. Recently, the sample size for GWAS has increased to tens of thousands or hundreds of thousands, and such large studies have newly discovered hundreds of variants involved in multiple common diseases [42, 49]. Most of these variants have very small effect sizes, emphatically supporting the message that the larger the association study the better it fares in discovering associations.

Unfortunately, some phenotypes are either logistically difficult or very expensive to collect. For these phenotypes, it is impractical to perform GWAS with tens of thousands or hundreds of thousands of individuals. Examples of these phenotypes include ones that require obtaining an inaccessible tissue such as brain expression, ones that require using a complex intervention such as a response to diet, and ones that require re-contacting individuals simply because they were unmeasured in the original cohort. For these phenotypes, investigators are often unable to collect samples large enough to discover variants with small effect sizes. As a result, it is unlikely that GWAS will be effectively conducted on these phenotypes.

One approach to increase power for GWAS on a phenotype that is hard to collect is utilizing an intermediate or proxy phenotype that is correlated to the target phenotype of interest. In this approach, one intermediate or proxy phenotype, which is highly correlated and easily collectable, is collected and then GWAS is performed on the intermediate phenotype in order to detect associated signals. For example, triglyceride levels can be collected as a proxy for obtaining information about metabolic diseases. This approach is known as intermediate phenotype analysis [16, 33].

One way to interpret the intermediate phenotype analysis is to consider the target phenotype as missing data and the use of intermediate phenotype as inferring the missing data. This connection

to missing data analysis motivates the following intuition. In missing data analyses, it is well known that utilzing multiple sources of information can be more effective than using a single source of information, which has been shown in machine learning [2, 15, 29, 39, 45] and genetics [5, 7, 48]. This motivates an intuition that utilizing multiple phenotypes together as proxies for a trait can lead to better performance, which is the basis of our approach.

In this paper, we propose an approach called *phenotype imputation* that allows one to perform GWAS on a phenotype that is difficult to collect. In our approach, we leverage the correlation structure between multiple phenotypes to impute the uncollected phenotype. Specifically, we estimate the correlation structure from a complete dataset that includes all phenotypes, and use the conditional distribution based on the multivariate normal (MVN) statistical framework to impute the uncollected phenotype in an incomplete dataset. Because our imputation approach utilizes only phenotypic information and not genetic information, imputed phenotypes can be subsequently used for association test without incurring data re-use. For the situations that the final GWAS will include both the complete and incomplete datasets, we provide an optimal meta-analysis strategy that combines association results from the collected phenotype and imputed phenotype while accounting for imputation uncertainties. Moreover, we demonstrate that we can analytically calculate the statistical power of association test using imputed phenotype, which can be helpful for study design purposes. In addition, we show that the summary statistic of the imputed phenotype can be approximated by a weighted linear combination of summary statistics for the proxy phenotypes. This result makes our method applicable to datasets where we only have access to the summary statistics and not the raw genotypes and phenotypes.

We show the effectiveness of our proposed approach by applying it to the Northern Finland Birth Cohort (NFBC) data [41]. By imputing the triglycerides (TG), body mass index (BMI), and systolic blood pressure (SBP) phenotypes, we recovered most of the significantly associated loci in the original data at the nominal significance level. This shows that even when the imputed phenotype may not provide sufficient power for discovery purposes due to imputation uncertainties, it can effectively be used for replication purposes. Our method is available at `http://genetics.cs.ucla.edu/phenIMP`.

2 Material and Methods

2.1 A Standard Genome-wide Association Study (GWAS)

We first describe the standard GWAS framework for testing genetic effects on quantitative phenotypes. Since the single nucleotide polymorphism (SNP) is the most common form of genetic variation, throughout this paper, we consider SNPs. However, the frameworks can be generalized to other types of variants. Suppose that we collect genotypes of m SNPs and ℓ quantitative phenotypes for n individuals. Let Y indicate a $(n \times \ell)$ matrix of phenotypic values where Y_k is a $(n \times 1)$ vector for the k-th phenotype. Let y_{jk} be the phenotypic value of the j-th individual for the k-th phenotype and $g_{ji} = \{0, 1, 2\}$ be the minor allele count of the j-th individual at the i-th SNP. Let p_i indicate the frequency of i-th variant in the population. In order to simplify the derivations, we standardize the minor allele counts for each SNP to have a mean zero and a variance one, such that $x_{ji} \in \{\frac{-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{1-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{2-2p_i}{\sqrt{2p_i(1-p_i)}}\}$ represents the standardized value of g_{ji}. Let X_i be the $(n \times 1)$ vector of standardized minor allele counts at the i-th SNP, where $\mathbf{1}^T X_i = 0$ and $X_i^T X_i = n$. We assume Fisher's polygenic model where the phenotype and the genotype follow normal distributions. Under the additive model that each SNP contributes linearly towards the phenotype:

$$Y_k = \mu_k \mathbf{1} + \sum_{i=1}^{m} \beta_{ik} X_i + \boldsymbol{e}_k \tag{1}$$

where μ_k is the phenotypic mean for the k-th phenotype, $\mathbf{1}$ is a $(n \times 1)$ vector of all ones, and β_{ik} is the effect of the i-th SNP towards the k-th phenotype. $\boldsymbol{e}_k \sim N(0, \sigma^2_{e_k} \boldsymbol{I})$ is the environment and measurement errors where $\boldsymbol{I}$ is an identity matrix. We additionally assume that the phenotypes are standardized so that their means are zero and their variances are one.

In a standard GWAS, we consider one SNP and one phenotype at a time. For notation clarity, we omit SNP index below (e.g. instead of X_i, we use X). The following model is used to test each SNP :

$$Y_k = \mu_k \mathbf{1} + \beta_k X + \boldsymbol{e}_k \tag{2}$$

Equation (2) is different from Equation (1) in that it omits the effects of the other SNPs, which can

manifest as background genetic effects. This was the motivation of using mixed model [24, 27, 28, 54] in the situations that sample data has population structures. Equation (2) leads us to least square solutions, $\hat{\mu}_k = \frac{\mathbf{1}^T X}{n}$ and $\hat{\beta}_k = \frac{X^T Y_k}{X^T X}$, where "hat" over parameters denotes estimated values. $\hat{e}_k = Y_k - \hat{\mu}\mathbf{1} - X\hat{\beta}_k$ is the residual error which is used to compute the standard error $\hat{\sigma}_k = \sqrt{\frac{\hat{\mathbf{e}}_k^T \hat{\mathbf{e}}_k}{n-2}}$ [17, 20, 21, 31]. Note that the estimated effect size is equal to the correlation between the standardized minor allele counts and the standardized phenotypic values, $\hat{\beta}_k = \text{cor}(X,Y_k)$. If the sample size is large enough, $\hat{\beta}_k$ follows a normal distribution with the mean equal to the true effect size β_k. Thus, we can define a normally-distributed association statistic as $s_k = \frac{\hat{\beta}_k \sqrt{n}}{\hat{\sigma}_k}$. Under the null hypothesis of no association ($\beta_k = 0$), the statistic s_k follows the standard normal distribution. Under the alternative hypothesis of true association, the statistic s_k follows a normal distribution with non-centrality parameter (NCP) $\lambda\sqrt{n} = \frac{\beta_k}{\sigma_k}\sqrt{n}$ [18, 20, 24, 54]:

$$s_k = \frac{\hat{\beta}_k}{\hat{\sigma}_k}\sqrt{n} \sim \begin{cases} N(0,1) & \text{null hypothesis (no association)} \\ N(\lambda\sqrt{n},1) & \text{alternative hypothesis} \end{cases} \tag{3}$$

To reject the null hypothesis of no association, given the significance threshold α, we compute the p-value, which is the probability that the observed statistic s_k will be more extreme under the null hypothesis, and determine that the association is significant if this probability is less than the significance threshold α (e.g. $\alpha = 5 \times 10^{-8}$ in GWAS). Equivalently, we reject the null hypothesis when $\Phi(s_k) < \alpha_s/2$ or $\Phi(s_k) > 1 - \alpha_s/2$, where $\Phi(.)$ indicates the cumulative density function of the standard normal distribution.

The statistical power is the probability of detecting an association under the situation that an association is present with a certain effect size [18, 35, 44, 46]. Intuitively, power measures the probability that the truly associated variants will be discovered. Since statistical power depends on both the effect size and the number of individuals in the study, power estimate can guide the choice of study size as well as providing expectations on what effect sizes can and can not be discovered. Given the effect size β_k, its standard error σ_k, the number of individuals n, and the significance threshold α, power is estimated as

$$P(\alpha, \beta_k, \sigma_k, n) = \Phi(\Phi^{-1}(\alpha/2) - \frac{\beta_k}{\sigma_k}\sqrt{n}) + 1 - \Phi(\Phi^{-1}(1-\alpha/2) - \frac{\beta_k}{\sigma_k}\sqrt{n}). \tag{4}$$

2.2 Phenotype Imputation

2.2.1 Phenotype Imputation Method

We consider two phenotype datasets in which we collected ℓ phenotypes from n_1 and n_2 individuals respectively. Let $\mathrm{Y}^{(1)}$ and $\mathrm{Y}^{(2)}$ be matrices of phenotypic values of size $(n_1 \times \ell)$ and $(n_2 \times \ell)$, and $Y_k^{(1)}$ and $Y_k^{(2)}$ be vectors of phenotypic values for the k-th phenotype in the first and second datasets respectively. We use $\neg\ell$ to indicate phenotypes excluding the ℓ-th phenotype. Thus, $y_{j\neg\ell}^{(1)}$ and $y_{j\neg\ell}^{(2)}$ are row vectors of the j-th individual phenotypes excluding the ℓ-th phenotype in $\mathrm{Y}^{(1)}$ and $\mathrm{Y}^{(2)}$ respectively.

We assume the phenotypic values follow a multivariate normal distribution. In the discussion section, we discuss the case where this assumption is violated. Assuming that each phenotype is standardized to mean zero and variance one, we model the joint distribution of multiple phenotypes as

$$\begin{bmatrix} y_{j1}^{(1)} \\ y_{j2}^{(1)} \\ \vdots \\ y_{j\ell}^{(1)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r_{12} & \cdots & r_{1\ell} \\ r_{21} & 1 & \cdots & r_{2\ell} \\ \vdots & & & \\ r_{(\ell-1)1} & r_{(\ell-1)2} & \cdots & r_{(\ell-1)\ell} \\ r_{\ell 1} & r_{\ell 2} & \cdots & 1 \end{bmatrix} \right).$$

We can represent this more compactly with a block matrix:

$$\begin{bmatrix} y_{j\neg\ell}^{(1)T} \\ y_{j\ell}^{(1)} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{\neg\ell} & R_{\neg\ell\ell} \\ R_{\neg\ell\ell}^T & 1 \end{bmatrix} \right) = \mathcal{N}\left(\mathbf{0}, H\right),$$

where $y_{j\neg\ell}^{(1)}$ is a row vector for the first $(\ell-1)$ phenotypic values for the j-th individual obtained from $\mathrm{Y}^{(1)}$ and $y_{j\neg\ell}^{(1)T}$ is the same vector in column format. Let $r_{k_1k_2}$ indicate the correlation between the two phenotypes k_1 and k_2, and let $R_{\neg\ell\ell} = [r_{1\ell}, r_{2\ell}, \cdots r_{\ell-1\ell}]^T$ denote a $((\ell-1)\times 1)$ vector of correlations between $Y_\ell^{(1)}$ and the phenotypes in $\mathrm{Y}^{(1)}$ excluding the ℓ-th phenotype. $\Sigma_{\neg\ell}$ is a $((\ell-1)\times(\ell-1))$ covariance matrix between the phenotypes in $\mathrm{Y}^{(1)}$ excluding the ℓ-th phenotype.

Using the above joint distribution, we condition on $y^{(1)}_{j\neg\ell}$ phenotypes to compute the distribution of phenotypic values for the j-th individual for the ℓ-th phenotype. This distribution is computed as follows:

$$(y^{(1)}_{j\ell} \mid y^{(1)}_{j\neg\ell}) \sim \mathcal{N}\left(R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}y^{(1)T}_{j\neg\ell}, 1 - R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}\right). \tag{5}$$

In the phenotype imputation problem, we assume that the ℓ-th phenotype is not collected in the second study. Let $\hat{Y}^{(2)}_\ell$ be the imputed phenotypic values for the uncollected phenotype. We assume the correlation between any pair of phenotypes is the same in two datasets $\mathrm{Y}^{(1)}$ and $\mathrm{Y}^{(2)}$. As a result, the above joint distribution in Equation (5) holds for $\mathrm{Y}^{(2)}$. Thus, we can perform similar conditional analysis. The conditional distribution is computed as follows:

$$(y^{(2)}_{j\ell} \mid y^{(2)}_{j\neg\ell}) \sim \mathcal{N}\left(R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}y^{(2)T}_{j\neg\ell}, 1 - R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}\right). \tag{6}$$

To impute the missing phenotype for a particular individual j, we use the mean of the conditional distribution as shown in Equation (6), $R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}y^{(2)T}_{j\neg\ell}$, as our prediction. A more compact formula to impute the ℓ-th phenotype for all the individuals in the dataset $\mathrm{Y}^{(2)}$ is as follows:

$$\hat{Y}_\ell = \mathrm{Y}^{(2)}_{\neg\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell} \tag{7}$$

Equation (7) shows that the imputed phenotype is a linear weighted combination of other collected phenotypes. Thus, if our multivariate normal assumption holds, the imputed phenotype will also follow a normal distribution.

Utilizing the imputed phenotype in the association study, we compute the association statistic of the imputed phenotype as the ratio between the estimated effect size for the imputed phenotype and its standard error. The association statistic is:

$$\hat{s}_\ell = \frac{\hat{\beta}'_\ell}{\hat{\sigma}'_\ell}\sqrt{n_2} = \frac{\frac{X^T\hat{Y}_\ell}{X^TX}}{\sqrt{\frac{\hat{\boldsymbol{e}}'^T_\ell\hat{\boldsymbol{e}}'_\ell}{n_2-2}}}\sqrt{n_2} \tag{8}$$

where $\hat{\beta}'_\ell$, $\hat{\sigma}'_\ell$, and $\hat{\boldsymbol{e}}'_\ell$ are estimated effect size, standard error, and residual error computed from

the imputed values of the ℓ-th phenotype respectively. Given a sufficiently large sample size, this statistic will follow a normal distribution. It will follow $\mathcal{N}(0,1)$ under the null hypothesis of no association to imputed phenotype.

2.2.2 Noisy Measurement Model

Here we introduce a model that is closely related to our phenotype imputation method. Under this model, called noisy measurement model (NMM), our method has interesting optimal properties that are related to the weighted sum of statistics approach. Note that, however, NMM is not a requirement for our method to work.

Under NMM, we assume that the phenotype ℓ has the main genetic effect, and other phenotypes can be modeled as the phenotype ℓ plus noise. That is, we consider the other phenotypes as noisy measurements of the phenotype ℓ. Under this model, obviously, the pleiotropic genetic effects to other phenotypes are driven by the main genetic effect to phenotype ℓ. As a result, the observed genetic effect to each of $\ell - 1$ phenotypes cannot be greater than the genetic effect to phenotype ℓ. This can be a strict assumption in general, but considering our situation that only phenotype ℓ is missing, this can be a reasonable assumption; if the genetic effect is greater in phenotype $k \neq \ell$, such that it makes more sense to model the main effect driven by phenotype k, analyzing the collected phenotype k data alone would be optimal, and we do not even need to perform phenotype imputation.

Specifically, we describe NMM as

$$Y_k^{(2)} = \frac{Y_\ell^{(2)} + \boldsymbol{u}_k}{\sqrt{1 + \sigma_{u_k}^2}} \tag{9}$$

where $\boldsymbol{u}_k$ is "noise" in the measurement. We assume that the noise follows a normal distribution with mean zero and variance $\sigma_{u_k}^2$, and further assume that the noise is independent from genotypes. The denominator was formulated to standardize the phenotype Y_k^2.

Let $r_{k\ell}$ be the correlation between $Y_\ell^{(2)}$ and $Y_k^{(2)}$. It is straightforward to show that,

$$r_{k\ell} = \sqrt{\frac{1}{1 + \sigma_{u_k}^2}}$$

Thus, we can re-write Equation (9) such as

$$Y_k^{(2)} = r_{k\ell}(Y_\ell^{(2)} + \boldsymbol{u}_k) \tag{10}$$

An important property of NMM is that if NMM holds, the strength of the effect of the variant on phenotype k is approximately the strength of the effect of the variant on phenotype ℓ times the correlation between the two phenotypes. That is, if $s_\ell \sim N(\lambda\sqrt{n_2}, 1)$, then approximately $s_k \sim N(r_{k\ell}\lambda\sqrt{n_2}, 1)$. This can be shown by,

$$\begin{aligned} s_k &= \frac{\frac{X^T Y_k^{(2)}}{X^T X}}{\sqrt{\frac{\hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{e}}_k}{n_2-2}}} r_{k\ell}\sqrt{n_2} = \frac{\frac{X^T Y_\ell^{(2)}}{X^T X}}{\sqrt{\frac{\hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{e}}_k}{n_2-2}}} r_{k\ell}\sqrt{n_2} + \frac{\frac{\boldsymbol{u}_k}{X^T X}}{\sqrt{\frac{\hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{e}}_k}{n_2-2}}} r_{k\ell}\sqrt{n_2} \\ &= r_{k\ell}\sqrt{\frac{\hat{\boldsymbol{e}}_\ell^T \hat{\boldsymbol{e}}_\ell}{\hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{e}}_k}} s_\ell + \frac{\frac{\boldsymbol{u}_k}{X^T X} r_{k\ell}}{\sqrt{\frac{\hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{e}}_k}{n_2-2}}}\sqrt{n_2} \\ s_k &\sim N(r_{\ell k}\lambda\sqrt{n_2}, 1) \end{aligned}$$

where we further assumed that the residual errors are similar for two phenotypes ($\hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{e}}_k \approx \hat{\boldsymbol{e}}_\ell^T \hat{\boldsymbol{e}}_\ell$), which holds true if the genetic effects are small. Note that a similar relationship arises when considering the statistics of two SNPs in linkage disequilibrium (LD) and the correlation between the two SNPs is r. It has been shown in various works [1, 12, 25, 38] the ratio between the NCPs of two statistics is the same as r. This is similar to NMM in the sense that a causal SNP drives the genetic effect, and the proxy SNP can be thought of as a noisy measurement of the causal SNP due to LD.

2.2.3 Power of Phenotype Imputation

If NMM describes truth, it is possible to analytically calculate the power of our phenotype imputation method. Under NMM, we consider the situation that the variant we are testing is associated with the ℓ-th phenotype with NCP of $\lambda\sqrt{n_2}$. As shown above, the NCP of the association statistic for the k-th phenotype on the same variant is $r_{k\ell}\lambda\sqrt{n_2}$ where $r_{k\ell}$ is the correlation between the phenotypes k and ℓ. Here, instead of considering the correlation between the phenotype ℓ and another phenotype k, we consider the correlation between the phenotype ℓ and the imputed phenotype of ℓ.

The covariance of the imputed and true phenotype is:

$$\mathrm{Cov}(\hat{Y}_\ell, Y_\ell) = \mathrm{Cov}(\mathrm{Y}^{(2)}_{\neg\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}, Y^{(2)}_\ell) = \mathrm{Cov}(\mathrm{Y}^{(2)}_{\neg\ell}, Y^{(2)}_\ell)\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell} = R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell} \tag{11}$$

We know that the variance of $Y^{(2)}_\ell$ is one, because we have already standardized the phenotypes. We compute the variance of the imputed phenotype as follows:

$$\begin{aligned}
\mathrm{Var}(\hat{Y}^{(2)}_\ell) &= \mathrm{Var}(\mathrm{Y}^{(2)}_{\neg\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}) \\
&= R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}\mathrm{Var}(\mathrm{Y}^{(2)}_{\neg\ell})\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell} \\
&= R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}\Sigma_{\neg\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell} \\
&= R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}
\end{aligned} \tag{12}$$

Utilizing the covariance between the imputed and true phenotype and the variance of phenotypes, we can compute the correlation as follows:

$$\mathrm{Cor}(\hat{Y}^{(2)}_\ell, Y^{(2)}_\ell) = \frac{\mathrm{Cov}(\hat{Y}^{(2)}_\ell, Y^{(2)}_\ell)}{\sqrt{\mathrm{Var}(\hat{Y}^{(2)}_\ell)}} = \sqrt{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}} \tag{13}$$

Under NMM, each phenotype is modeled as a standardized linear combination of phenotype ℓ and noise. Since imputed phenotype is also a linear combination of those phenotypes, we can consider the imputed phenotype as a new phenotype that we can apply NMM. That is, we can consider the imputed phenotype as a noisy version of the true phenotype. Then, by the property of NMM,

$$\begin{aligned}
\mathrm{Cov}(\hat{s}_\ell, s_\ell) &= \sqrt{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}} = r_{imp} \\
\hat{s}_\ell &\sim \mathcal{N}(\sqrt{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}}\lambda\sqrt{n_2}, 1)
\end{aligned} \tag{14}$$

Since we obtained NCP of the statistic for imputed phenotype, we can analytically calculate power of our phenotype imputation using Equation (4).

Note that the following quantity will have mean zero,

$$\hat{s}_\ell - r_{imp}s_\ell \sim N(0, 1 - r^2_{imp}) \tag{15}$$

The variance of $\hat{s}_\ell - r_{imp} s_\ell$ is computed as follow:

$$\begin{aligned} \mathrm{Var}(\hat{s}_\ell - r_{imp} s_\ell) &= \mathrm{Var}(\hat{s}_\ell) + r_{imp}^2 \mathrm{Var}(s_\ell) - 2 r_{imp} \mathrm{Cov}(\hat{s}_\ell, s_\ell) \\ &= 1 + r_{imp}^2 - 2 r_{imp}^2 = 1 - r_{imp}^2 \end{aligned}$$

In Results, we evaluate this quantity in real dataset, to evaluate if our imputation method works as expected.

2.2.4 Relation to Optimal Linear Combinations of Marginal Statistics

The result of phenotype imputation is a weighted linear combination of the observed phenotypes. Here, we show that under NMM, phenotype imputation is the "optimal" weighted combination of the phenotypes in terms of statistical power. Let $S_{\neg\ell}$ be a vector of association statistics computed for the first $\ell - 1$ phenotypes, $S_{\neg\ell} = [s_1, s_2, \cdots s_{\ell-1}]^T$. Under NMM, given that the NCP of the uncollected phenotype is $\lambda\sqrt{n_2}$, we have $S_{\neg\ell} \sim N(R_{\neg\ell\ell}\lambda\sqrt{n_2}, \Sigma_{\neg\ell})$. We calculate the association statistic of the imputed phenotype as a linear combination of weighted statistics computed for the $(\ell-1)$ phenotypes. Let $W = \{w_1, w_2, \cdots w_{\ell-1}\}$ indicate the vector of weights where w_i is the weight corresponding to the i-th phenotype marginal statistics. Thus, we have:

$$W^T S_{\neg\ell} \sim N(W^T R_{\neg\ell\ell}\lambda\sqrt{n_2}, W^T \Sigma_{\neg\ell} W) \tag{16}$$

Using the above formula and the fact the variance of the associated statistic is one, we have:

$$\hat{s}_\ell \sim \mathcal{N}(\frac{W^T R_{\neg\ell\ell}}{\sqrt{W^T \Sigma_{\neg\ell} W}}\lambda\sqrt{n_2}, 1)$$

It has been shown, power is maximized when we maximize the NCP [13]. Thus, we find the set of weights which maximizes $\frac{W^T R_{\neg\ell\ell}}{\sqrt{W^T \Sigma_{\neg\ell} W}}$. Let $A^T A = \Sigma_{\neg\ell}$ and $W' = AW$, our maximization problem reduces to following optimization:

$$\underset{W'}{\mathrm{argmax}} \frac{W'^T A \Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}}{\sqrt{W'^T W'}}.$$

Let $\Theta = A\Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}$. Using the Cauchy-Schwarz inequality, we have

$$\sum_{j=1}^{\ell-1} w'_j \theta_j \leq \sqrt{\sum_{j=1}^{\ell-1} w'^2_j} \sqrt{\sum_{j=1}^{\ell-1} \theta_j^2}.$$

The optimal value for W' is Θ and the maximum NCP is as follows:

$$\sqrt{R_{\neg\ell\ell}^T \Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}} \lambda \sqrt{n_2}.$$

This is exactly the NCP obtained from the previous section. Moreover, the optimal value for W is $\Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}$ which is the same vector of weights used in the previous section. This is the justification for Equation (14) above.

Interestingly, this result indicates we can use Equation (16) and the optimal weights, which are obtained in this section, to estimate the marginal statistics of the imputed phenotype as weighted linear combinations of observed marginal statistics from other phenotypes. Thus, given the observed marginal statistics of the first $(\ell - 1)$ phenotypes and the pairwise phenotype correlations, we can compute the estimated marginal statistics. Our method does not need to have access the raw genotypes and phenotypes. This makes our method applicable to datasets for which we only have access to the summary statistics.

We note that for any vector of weights, including the ones utilized in imputation, the type I error rates are controlled. This is because if the variant we are testing is not associated with the phenotype, $\lambda = 0$, then the NCP of the imputed statistic for that variant is zero.

2.2.5 Optimal Meta-Analysis Strategy for Combining Imputed and Observed Values

We use the phenotype imputation to fill the values of the phenotype for individuals whose phenotypic values are missing. We then want to obtain an association statistic for the combined dataset including the imputed and observed phenotypes. However, since our imputation is not always accurate, utilizing both observed and imputed data together without distinguishing them is suboptimal. We propose to compute the association statistics by performing statistical tests on the collected phenotype and imputed phenotype separately. Then, we perform a fixed-effect meta-analysis to combine the two statistics.

We use Y_m and Y_c to indicate the missing and collected phenotypes respectively. We compute the association statistic of each set separately. The association statistic for the collected phenotype is computed as $s_c \sim \mathcal{N}(\lambda_c\sqrt{n_c}, 1)$ where λ_c is the NCP of the phenotype and n_c is the number of individuals whose phenotypic values are collected for this phenotype. We use Equation (14) to compute the z-score for the imputed phenotype as $\hat{s}_m \sim \mathcal{N}(\sqrt{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}}\lambda_c\sqrt{n_m}, 1)$ where n_m is the number of individuals whose phenotypic values are missing for this phenotype.

We combine the two statistics using the fixed-effects meta-analysis. The fixed-effects meta-analysis association statistic, s_{FE}, is computed as $s_{FE} = \frac{w_c s_c + w_m \hat{s}_m}{\sqrt{w_c^2 + w_m^2}}$, where w_c and w_m are computed such that the meta-analysis association statistic is maximized [11, 50]. As shown in previous studies [11, 51] the optimal weights are computed as $w_c = \sqrt{n_c}$ and $w_m = \sqrt{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}n_m}$. Thus, we have:

$$s_{FE} = \frac{\sqrt{n_c}s_c + \sqrt{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}n_m}\hat{s}_m}{\sqrt{n_c + R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}R_{\neg\ell\ell}n_m}} \tag{17}$$

Using Equation (17), we combine the statistics computed for the collected phenotype and imputed phenotype as a joint association statistic.

2.2.6 Polygenic Model

We described the properties of our method under NMM. However, NMM is a simple model and may not always hold true. Here we introduce a more complex model, which explicitly models both the genetic and environmental correlations in phenotypes. We suggest a strategy that is optimized for this model, and we show that the new strategy is equivalent to our standard strategy under some simplifying assumptions.

Let $B = \{\beta_1, \beta_2, \cdots \beta_\ell\}$ indicate the vector of true effect sizes of a given variant towards all ℓ phenotypes where β_j is the effect size for the j-th phenotype. Let E be a $(n \times \ell)$ matrix which models the errors. We consider a multi-phenotype setting, where we perform a join testing of a variant for all the ℓ phenotypes:

$$\text{vec(Y)} = (\boldsymbol{I} \otimes X)B + \text{vec}(E)$$

where vec() is an operator that converts a matrix to vector by stacking columns of matrix and $\otimes$ is an operator that performs Kronecker product between two matrices.

Given this multi-phenotype setting, we can model the genetic and environmental correlations. Let ρ_{ij} and ξ_{ij} indicate the genetic and environment correlations, respectively, between i-th and j-th phenotype. Let σ^2_{gi} denotes the genetic variance of phenotype i. Let σ^2_{ei} denotes the error variance of phenotype i. In the multi-phenotype polygenic model, the true vector of effect sizes are assumed to follow a MVN [14, 52–54], such that

$$\begin{bmatrix} \beta_1^{(1)} \\ \beta_2^{(1)} \\ \vdots \\ \beta_\ell^{(1)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \frac{1}{m} \begin{bmatrix} \sigma^2_{g_1} & \rho_{12}\sigma_{g_1}\sigma_{g_2} & \cdots & \rho_{1\ell}\sigma_{g_1}\sigma_{g_\ell} \\ \rho_{21}\sigma_{g_1}\sigma_{g_2} & \sigma^2_{g_2} & \cdots & \rho_{2\ell}\sigma_{g_2}\sigma_{g_\ell} \\ \vdots & & & \\ \rho_{\ell 1}\sigma_{g_\ell}\sigma_{g_1} & \rho_{\ell 2}\sigma_{g_\ell}\sigma_{g_2} & \cdots & \sigma^2_{g_\ell} \end{bmatrix} \right) = \mathcal{N}(\mathbf{0}, \frac{1}{m}G) \tag{18}$$

where $\frac{1}{m}$ is the proportion that the variant contributes to the genetic variance [14, 52–54]. Here, we assumed that $\frac{1}{m}$ is the same for all phenotypes. In the similar way, we define a $(\ell \times \ell)$ variance matrix that encodes the environmental correlations,

$$\Upsilon = \begin{bmatrix} \sigma^2_{e_1} & \xi_{12}\sigma_{e_1}\sigma_{e_2} & \cdots & \xi_{1\ell}\sigma_{e_1}\sigma_{e_\ell} \\ \xi_{21}\sigma_{e_1}\sigma_{e_2} & \sigma^2_{e_2} & \cdots & \xi_{2\ell}\sigma_{e_2}\sigma_{e_\ell} \\ \vdots & & & \\ \xi_{\ell 1}\sigma_{e_\ell}\sigma_{e_1} & \xi_{\ell 2}\sigma_{e_\ell}\sigma_{e_2} & \cdots & \sigma^2_{e_\ell} \end{bmatrix}$$

Under the polygenic model, we have $\text{Cov}(Y_i, Y_i) = \sigma^2_{g_i}K + \sigma^2_{e_i}\boldsymbol{I}$ and $\text{Cov}(Y_i, Y_j) = \rho_{ij}\sigma_{g_i}\sigma_{g_i}K + \xi_{ij}\sigma_{e_i}\sigma_{e_j}\boldsymbol{I}$ where K is the kinship matrix that represents the genetic relatedness between individuals. We use a $(\ell n \times \ell n)$ matrix that encodes the covariance for all pairs of phenotypes and let V represent this covariance matrix:

$$V = \begin{bmatrix} \text{Cov}(Y_1, Y_1) & \text{Cov}(Y_1, Y_2) \cdots \text{Cov}(Y_1, Y_\ell) \\ \text{Cov}(Y_2, Y_1) & \text{Cov}(Y_2, Y_2) \cdots \text{Cov}(Y_2, Y_\ell) \\ \vdots & \\ \text{Cov}(Y_\ell, Y_1) & \text{Cov}(Y_\ell, Y_2) \cdots \text{Cov}(Y_\ell, Y_\ell) \end{bmatrix} = G \otimes K + \Upsilon \otimes \boldsymbol{I}$$

Let $\hat{B}$ indicate the vector of estimated effect sizes for all the ℓ phenotypes for a given variant. Using the mixed model we have $\hat{B} = ((\boldsymbol{I} \otimes X)^T V^{-1} (\boldsymbol{I} \otimes X))^{-1} (\boldsymbol{I} \otimes X)^T V^{-1} Y$ and $\text{Var}(\hat{B}) = ((\boldsymbol{I} \otimes X)^T V^{-1} (\boldsymbol{I} \otimes X))^{-1} = \Psi$. Let ψ_{ij} be the ith row and jth column element of Ψ. We can compute the joint distribution of marginal statistics for all the ℓ phenotypes. Let $S = \{s_1, s_2, \cdots s_\ell\}$ indicate a $(\ell \times 1)$ vector of marginal statistics. The joint distribution of statistics follows a MVN which is as follows:

$$S \sim \mathcal{N}\left(\begin{bmatrix} \frac{\beta_1}{\psi_{11}} \\ \vdots \\ \frac{\beta_\ell}{\psi_{\ell\ell}} \end{bmatrix}, \begin{bmatrix} 1 & \frac{\psi_{12}}{\sqrt{\psi_{11}\psi_{22}}} & \cdots \frac{\psi_{1\ell}}{\sqrt{\psi_{11}\psi_{\ell\ell}}} \\ \frac{\psi_{21}}{\sqrt{\psi_{11}\psi_{22}}} & 1 & \cdots \frac{\psi_{2\ell}}{\sqrt{\psi_{22}\psi_{\ell\ell}}} \\ \vdots & & \\ \frac{\psi_{\ell 1}}{\sqrt{\psi_{\ell\ell}\psi_{11}}} & \frac{\psi_{2\ell}}{\sqrt{\psi_{\ell\ell}\psi_{22}}} & \cdots 1 \end{bmatrix} \right) = \mathcal{N}(\Lambda, \Gamma) \tag{19}$$

where Λ is the vector of NCPs. Note that using Equation (18), we can assume a prior distribution for effect size of the single SNP that we test, such as $B \sim \mathcal{N}(0, \frac{1}{m} G)$. Since NCP is true effect size normalized to have variance one, prior distribution for B gives us prior distribution for NCP,

$$\Lambda \sim \mathcal{N}\left(0, \frac{1}{m} \begin{bmatrix} \frac{\sigma^2_{g_1}}{\psi_{11}} & \frac{\rho_{12}\sigma_{g_1}\sigma_{g_2}}{\sqrt{\psi_{11}\psi_{22}}} & \cdots \frac{\rho_{1\ell}\sigma_{g_1}\sigma_{g_\ell}}{\sqrt{\psi_{11}\psi_{\ell\ell}}} \\ \frac{\rho_{21}\sigma_{g_2}\sigma_{g_1}}{\sqrt{\psi_{11}\psi_{22}}} & \frac{\sigma^2_{g_2}}{\psi_{22}} & \cdots \frac{\rho_{2\ell}\sigma_{g_2}\sigma_{g_\ell}}{\sqrt{\psi_{22}\psi_{\ell\ell}}} \\ \vdots & & \\ \frac{\rho_{\ell 1}\sigma_{g_\ell}\sigma_{g_1}}{\sqrt{\psi_{\ell\ell}\psi_{11}}} & \frac{\rho_{\ell 2}\sigma_{g_\ell}\sigma_{g_2}}{\sqrt{\psi_{\ell\ell}\psi_{22}}} & \cdots \frac{\sigma^2_{g_\ell}}{\psi_{\ell\ell}} \end{bmatrix} \right) = \mathcal{N}(0, \Omega) \tag{20}$$

In summary, we have $S \sim \mathcal{N}(\Lambda, \Gamma)$ and $\Lambda \sim \mathcal{N}(0, \Omega)$. We assume the NCP for the ℓ-th phenotype

is $\lambda\sqrt{n_2}$. Thus, the NCPs of the phenotypes excluding, the ℓ-th phenotype is as follows:

$$\Lambda_{\neg\ell} \sim \mathcal{N}(\Omega_{\neg\ell\ell}^T \Omega_{\ell\ell}^{-1} \lambda\sqrt{n_2}, \Omega_{\neg\ell\neg\ell} - \Omega_{\neg\ell\ell}\Omega_{\ell\ell}^{-1}\Omega_{\neg\ell\ell}^T) \tag{21}$$

In similar way, the marginal statistics of all the phenotypes excluding the ℓ-th phenotype is as follows:

$$S_{\neg\ell} \sim \mathcal{N}(\Omega_{\neg\ell\ell}^T \Omega_{\ell\ell}^{-1} \lambda\sqrt{n_2}, \Omega_{\neg\ell\neg\ell} - \Omega_{\neg\ell\ell}\Omega_{\ell\ell}^{-1}\Omega_{\neg\ell\ell}^T + \Gamma_{\neg\ell\neg\ell}) \tag{22}$$

To simplify above equation, we can set the $\Lambda_{\neg\ell}$ to the mean of Equation (21). This assumption implies that the marginal statistics of all the phenotypes excluding, the ℓ-th phenotype is as follows:

$$S_{\neg\ell} \sim \mathcal{N}(\Omega_{\neg\ell\ell}^T \Omega_{\ell\ell}^{-1} \lambda\sqrt{n_2}, \Gamma_{\neg\ell\neg\ell}) \tag{23}$$

Similar to previous section, we consider the imputed marginal statistics is a weighted linear combination of all the marginal statistics that maximizes the power. Using Cauchy-Schwartz inequality, we can show that the maximum NCP of $\hat{s}_\ell$ will be $\sqrt{\Omega_{\ell\ell}^{-1}\Omega_{\neg\ell\ell}^T\Gamma_{\neg\ell\neg\ell}^{-1}\Omega_{\neg\ell\ell}\Omega_{\ell\ell}^{-1}}\lambda\sqrt{n_2}$. The maximum NCP is achieved when the weights of the marginal statistics are $\Gamma_{\neg\ell\neg\ell}^{-1}\Omega_{\neg\ell\ell}\Omega_{\ell\ell}^{-1}$. Therefore, we have successfully derived the weighted combination of marginal statistics that is optimized for the polygenic model.

2.2.7 Relation Between Polygenic Model and Noisy Measurement Model

We show that under some simplifying assumptions, the method for polygenic model is equivalent to the standard method for NMM. We make two assumptions that the pairwise genetic and environment correlations are equal (e.g. $\rho_{ij} = \xi_{ij}$) and that the individuals are sufficiently unrelated that we can approximate K with $\boldsymbol{I}$. The second assumptions implies that we have no population structure. Given these two assumptions, we can simplify V which is as follows:

$$V = \begin{bmatrix} (\sigma_{g_1}^2 + \sigma_{e_1}^2)\boldsymbol{I} & (\rho_{12}\sigma_{g_1}\sigma_{g_2} + \xi_{12}\sigma_{e_1}\sigma_{e_2})\boldsymbol{I} & \cdots (\rho_{1\ell}\sigma_{g_1}\sigma_{g_\ell} + \xi_{1\ell}\sigma_{e_1}\sigma_{e_\ell})\boldsymbol{I} \\ (\rho_{21}\sigma_{g_2}\sigma_{g_1} + \xi_{21}\sigma_{e_2}\sigma_{e_1})\boldsymbol{I} & (\sigma_{g_2}^2 + \sigma_{e_2}^2)\boldsymbol{I} & \cdots (\rho_{2\ell}\sigma_{e_2}\sigma_{g_\ell} + \xi_{2\ell}\sigma_{e_2}\sigma_{e_\ell})\boldsymbol{I} \\ \vdots & & \\ (\rho_{\ell 1}\sigma_{g_\ell}\sigma_{g_1} + \xi_{\ell 1}\sigma_{e_\ell}\sigma_{e_1})\boldsymbol{I} & (\rho_{\ell 2}\sigma_{\ell_1}\sigma_{g_2} + \sigma_{e_\ell}\sigma_{e_2})\boldsymbol{I} & \cdots (\sigma_{g_\ell}^2 + \sigma_{e_\ell}^2)\boldsymbol{I} \end{bmatrix} = H \otimes \boldsymbol{I} \tag{24}$$

where $\sigma_{g_i}^2 + \sigma_{e_i}^2 = 1$ for any phenotypes as we standardized the phenotypes and H is phenotypic correlation matrix. Thus, $\text{Var}(\hat{B}) = \left((\boldsymbol{I} \otimes X)^T (H \otimes \boldsymbol{I})^{-1} (\boldsymbol{I} \otimes X)\right)^{-1} = \frac{1}{n_2} H$. As a result, we have $\Lambda \sim \mathcal{N}\left(0, \frac{1}{mn_2} H\right)$. Given the NCP for the ℓ-th phenotype is $\lambda\sqrt{n_2}$ then the NCPs of all the phenotype excluding the ℓ-th phenotype will have a distribution with mean equal to $R_{\neg\ell\ell}\lambda\sqrt{n_2}$. Similar to previous section, if we fix NCP to its mean value for simplification, the method converges to the standard approach based on NMM. This result implies that considering the two assumptions mentioned above, our approach for the multi-phenotype polygenic model is equivalent to the standard strategy for NMM.

2.2.8 Avoiding Over-fitting

In some datasets, such as eQTL datasets, the number of phenotypes is large (ℓ is large). Thus, we have the risk of over-fitting. Over-fitting occurs in a method where the number of parameters is large. As the method usually does not generalize, it produces very high accuracy in the training dataset and very low accuracy in the test dataset. One way to avoid over-fitting, we can add a sparsity prior such as the Laplace prior [37] which reduces the linear regression to LASSO [47]. In the LASSO setting, we impute the phenotype while utilizing few phenotypes to avoid over-fitting. Another solution is to select the most informative phenotypes and then apply our method. As an example, we can pick the top 10 phenotypes based on their correlation with the target phenotype. We only use these 10 phenotypes in our method.

2.2.9 Handling Missing Data

Our method can handle missing data in the target dataset by performing imputation with only the available phenotypes for each individual. In this scenario, some of the individuals will have more accurate imputation than others because they utilize more phenotypes to perform the imputation. We have developed an optimal approach for performing association test utilizing these differing degrees of quality of phenotype imputation, which we detail in the Supplementary Materials.

2.2.10 Utilizing Covariates

In a typical GWAS, we usually adjust for the non-genetic factors that influence the phenotype, such as sex, age, study design, and known clinical covariates. Covariate adjustment reduces the spurious association signals in a study. Given, we have p covariates, we need to adjust for them by extending the Equation (1). Thus, the polygenic model used to handle covariates for the k-th phenotype is as follows:

$$Y_k = \mu_k \mathbf{1} + \sum_{i=1}^{m} \beta_{ik} X_i + \sum_{i=1}^{p} \gamma_{ik} Z_i + \boldsymbol{e}_k \tag{25}$$

where Z_i is the i-th covariate and γ_{ik} is the effect of that covariate towards the k-th phenotype. Moreover, to perform the single SNP association test instead of using Equation (2), we need to adjust for the covariates. We use the following model for the single SNP association test:

$$Y_k = \mu_k \mathbf{1} + \beta_k X + \sum_{i=1}^{p} \gamma_{ik} Z_i + \boldsymbol{e}_k \tag{26}$$

There are two possible ways to adjust for covariates for phenotype imputation. First possible way is to impute the phenotype and then use Equation (26) for asscoation testing. This testing is similar for testing collected phenotypes and adjusting for covariates. Second possible way is to regress out the covariates from all the collected phenotypes to generate new phenotypes where the covariates are removed. Then, we use our imputation method to impute the uncollected phenotype using the phenotypes which the covariates are regressed out. In this case, we can use Equation (2) to perform

asscoation testing.

3 Results

3.1 Overview of Phenotype Imputation

In phenotype imputation, we consider two datasets (D_1, D_2), in which multiple phenotypes are collected along with genetic information to perform GWAS. In the first dataset (D_1), we collect the target phenotype and the related phenotypes. In the second dataset (D_2), the related phenotypes have been collected for all of the individuals but the target phenotype has not been collected. Given these datasets, we predict the uncollected target phenotype in the second dataset (D_2) by leveraging the correlation structure between the additional phenotypes and the target phenotype. We use the first dataset (D_1) to approximate this correlation structure. After imputing the target phenotype, we perform GWAS to discover genetic variants that are significantly associated with the imputed target phenotype.

Our framework allows for the estimation of the relative power of imputation compared to the power if the phenotype was collected in the sample. Intuitively, the power loss depends on how close the imputed phenotypes are to the true phenotypes. We define the correlation between the imputed and true phenotypes as r_{imp} and we can estimate r_{imp} from only the first dataset. This allows us to have an idea of how well the imputation will perform in the target dataset. Under some additional assumptions, which we refer to as the noisy measurement model (NMM), the power in the imputed study with n individuals is equivalent to the power of a complete study where $r_{imp}^2 N$ individuals were collected (see Methods for the detailed derivation). We define the number of individuals that contribute toward the power of a statistical test for a phenotype as the effective number of individuals. For example, we can impute triglycerides (TG) levels in the NFBC dataset [41] using high-density lipoproteins (HDL), low-density lipoproteins (LDL), and systolic blood pressure (SBP) with a correlation of 0.5. As a result, in a study where we collect HDL, LDL, and SBP for 8,000 individuals, the power of GWAS on the imputed TG is equivalent to performing GWAS in 2,000 individuals where TG has been collected.

Phenotype	rsID	Real test data [1]				Imputed test data				$\lvert Z_{imp}\text{-}r_{imp}{}^{*}Z_{real}\rvert$
		β	se(β)	Z-score (Z_{real})	P-value	β	se(β)	Z-score (Z_{imp})	P-value	
TG	rs3923037	0.074	0.0149	4.96	7.14e-07	0.0224	0.0083	2.700	0.006	0.17
	rs6728178	0.076	0.0149	5.10	3.45e-07	0.0267	0.0083	3.209	0.001	0.24
	rs6754295	0.074	0.0149	4.94	7.91e-07	0.0266	0.0083	3.197	0.001	0.32
	rs676210	0.0752	0.0149	5.01	5.38e-07	0.0250	0.0083	2.996	0.002	0.084
	rs673548	0.0762	0.0149	5.08	3.81e-07	0.02530	0.0083	3.031	0.002	0.08
	rs1260326	-0.0807	0.0150	-5.37	8.15e-08	-0.004	0.0084	-0.534	0.59	2.58
	rs10096633	0.0819	0.0147	5.55	3.00e-08	0.0191	0.0082	2.324	0.02	0.79
BMI	rs987237	-0.074	0.0150	-4.97	6.63e-07	-0.037	0.00929	-4.07	4.62e-05	0.93
	rs11759809	-0.074	0.0150	-4.95	7.35e-07	-0.036	0.00931	-3.96	7.43e-05	0.84
SBP	rs782586	0.074	0.0149	4.96	7.43E-07	0.036	0.01016	3.50	0.00047	0.37
	rs782588	0.074	0.0149	4.94	8.14E-07	0.035	0.01014	3.43	0.00061	0.32
	rs782602	0.075	0.0150	5.01	5.53E-07	0.034	0.01016	3.39	0.00071	0.23
	rs2627759	0.070	0.0150	4.65	3.44E-06	0.032	0.01016	3.12	0.00183	0.19
	rs10486523	-0.073	0.0145	-4.98	6.62E-07	-0.031	0.00999	-3.08	0.00207	0.06
	rs9791555	-0.073	0.0145	-4.97	6.79E-07	-0.031	0.00999	-3.07	0.00214	0.06
	rs7799346	-0.073	0.0145	-4.98	6.52E-07	-0.030	0.00999	-3.04	0.00235	0.09
	rs6976779	0.069	0.0146	4.71	2.59E-06	0.039	0.01000	3.94	0.00008	0.97
	rs2846572	-0.067	0.0145	-4.62	3.94E-06	-0.031	0.00998	-3.10	0.00194	0.19

Table 1: Comparison between the association test on the real test data for TG, BMI, and SBP phenotypes and the imputed test data in the NFBC data. Z_{imp} and Z_{real} are the test statistics (Z-score) obtained from the imputed and original datasets respectively. The last column is the difference between the imputed test statistics and the analytical test statistics.

3.2 Phenotype Imputation Controls Type I Error

We simulated datasets for multiple phenotypes under the null model where the variant we are testing has no effect (effect size of zero) towards the target phenotype. We computed the type I error under five different significance thresholds: 0.05, 0.01, 0.005, 5×10^{-6}, and 5×10^{-8}. We generated 100,000,000 simulated datasets which consist of 1000 individuals. The type I error rates for our imputation method were 0.049, 0.0099, 0.00489, 4.90×10^{-6}, and 4.89×10^{-8} for the significance thresholds of 0.05, 0.01, 0.005, 5×10^{-6}, and 5×10^{-8}, respectively. This indicates that the type I error is correctly controlled in our imputation method. Using the Northern Finland Birth Cohort dataset [41] we show that the type I error is controlled (see Figure S.1). We plot the Q-Qplot of the z-score for the imputed triglycerides (TG) phenotype from the Finland dataset. There is no inflation in the Q-Qplot shown in Figure S.1

[1]The real test data is obtained from the NFBC data by removing the 500 individuals who are assumed to be missing in our experiment.

3.3 Phenotype Imputation on Northern Finland Birth Cohort (NFBC)

In order to assess the performance of our method, we utilize the Northern Finland Birth Cohort (NFBC) dataset [41]. The NFBC dataset consists of 10 phenotypes collected from 5,327 individuals. The 10 phenotypes are triglycerides (TG), high-density lipoproteins (HDL), low-density lipoproteins (LDL), glucose (GLU), insulin (INS), body mass index (BMI), C-reactive protein (CRP) as a measure of inflammation, systolic blood pressure (SBP), diastolic blood pressure (DBP), and height. The genotype data consists of 331,476 SNPs. Figure 1 shows the pairwise correlations between each pair of phenotypes. The correlation coefficients between the phenotypes in this data are between 0.01-0.62. SBP and DBP are the two phenotypes that show the highest correlation.

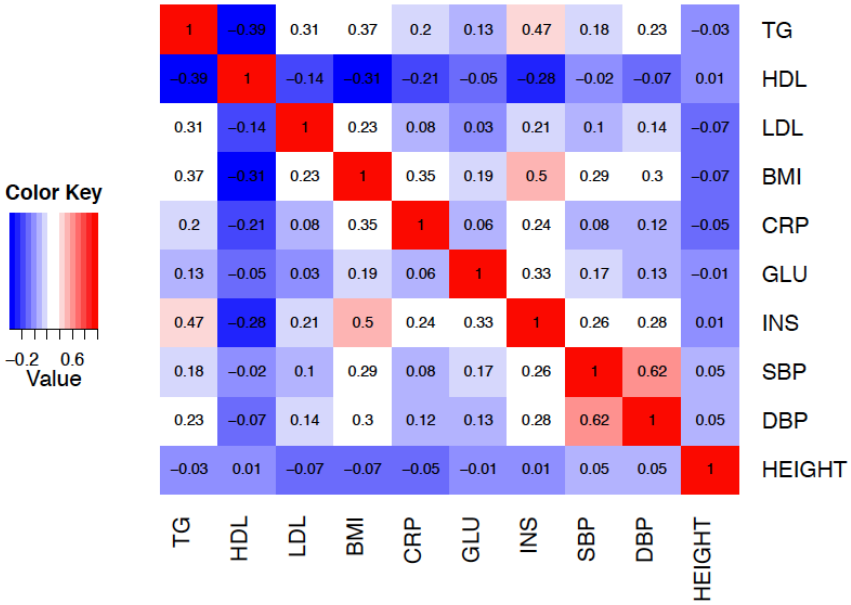

Figure 1: The pairwise correlation between each pair of phenotype in the NFBC dataset.

We consider the possibility of imputing each of these 10 phenotypes using the other nine phenotypes. We first compute the corresponding value of r_{imp} (Table S.1). In order to evaluate our method, we are interested in the scenario where r_{imp} is high and higher than the highest pairwise correlation. Among these 10 phenotypes the TG, INS, DBP, BMI and SBP are the phenotypes that satisfy these criteria. Since INS and DBP have no significantly associated variants, we focus on TG, BMI, and SBP phenotypes for our evaluation.

For our experiments, we assume that TG, BMI and SBP phenotypes are only collected for

500 individuals which are used as a training dataset to estimate the correlation structure between phenotypes. We mask the TG, BMI and SBP phenotypic values in the rest of the individuals and only use them when we measure the imputation accuracy. We utilize the 500 individuals to compute the correlation structure between the phenotypes, and we use our method to impute the TG, BMI and SBP phenotypes for the other individuals.

The correlation between the imputed phenotype and the true TG phenotypes is $r_{imp} = 0.58$. Our estimate of this correlation from the training data is $\hat{r}_{imp} = 0.58$. This correlation coefficient and the size of the data results in an effective number of individuals being $\sim$1620 ($0.58^2 \times (5,327 - 500) = 1623$). For this reason, we do not expect to see any significant loci in our imputed data. However, this size of data is enough to observe an effect in the context of a replication study. We perform association analysis using EMMAX [24] in the imputed phenotypes, and also utilizing the original TG phenotypes for comparison. Table 1 shows the estimated effect size(β), standard error of the estimated effect size ($se(\beta)$), Z-scores, and p-values. The result in Table 1 indicates that when we run EMMAX [24] on the original TG phenotype in the test dataset, we have seven loci that pass our significance threshold of 5×10^{-6}. When we run EMMAX [24] on the imputed phenotypes for these seven loci, we observe that most of these loci (six out of seven loci) pass the replication significance threshold of at least 0.05. Therefore, it appears that for most variants, phenotype imputation power is equivalent to collecting $r_{imp}^2 n$ individuals. Surprisingly, the test statistic (Z-score) for the imputed phenotype of all variants other than rs1260326 is close to r_{imp} times the test statistic (Z-score) at the actual variant as shown in the last column of Table 1. We define two statistics are close when the difference between the two statistics is less than one standard deviation (the standard deviation is 1). This is exactly the result we expect under NMM. We also expect that if the assumption holds, the distribution of the statistic on the imputed data minus r_{imp} times the statistic on the original data (last column of Table 1) over the whole data will follow a distribution with mean 0 and variance $1 - r_{imp}^2$ as described in the methods. In Figures 2, S.3, and S.4, we show that this is the case for the TG, BMI and SBP phenotypes respectively. This shows that although NMM is a simple model, NMM describes these datasets effectively. This result shows that performing GWAS on the imputed phenotype has enough power to identify most of the associated loci that are significant when we perform GWAS on the original phenotype.

We further investigate rs1260326 whose imputed z-score was not close to the expected value.

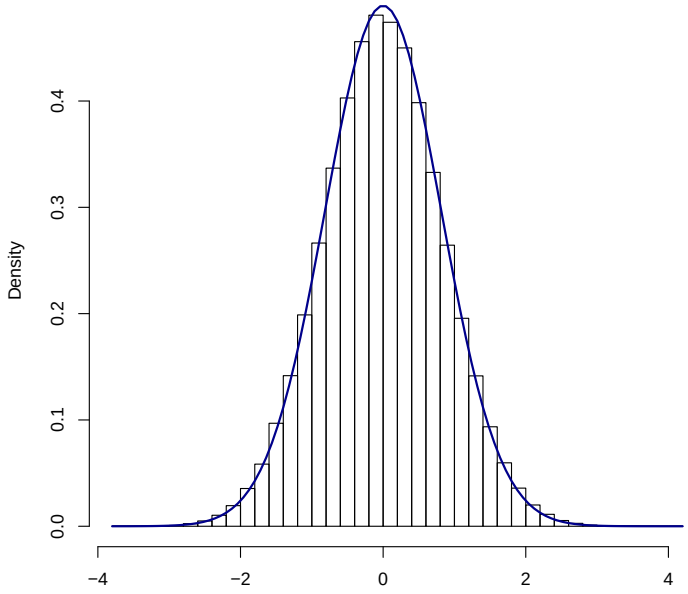

Figure 2: Difference between the imputed marginal statistics and analytical marginal statistics for TG phenotype. Imputed marginal statistics is obtained from the association between the genotype and the imputed phenotype and analytical marginal statistics is equal to the marginal statistics computed on the true target phenotype scaled by r_{imp}. The blue curve is the normal distribution with mean zero and variance 1-r^2_{imp}. This histogram indicates this difference follows a normal distribution (mean zero and variance 1-r^2_{imp}). Thus, for most null variants the NMM assumption holds.

Table S.2 shows the EMMAX [24] results for rs1260326 on all of the phenotypes in the NFBC data. We observe that in the original data this SNP is only significant for the TG phenotype. Thus, the effect sizes of this SNP for multiple phenotypes are not well modeled by the overall phenotypic correlation. For this reason, our method and any other possible approaches that will use proxy phenotypes, will have limited performance in detecting such a locus.

3.4 Phenotype Imputation on Hybrid Mouse Diversity Panel (HMDP)

We also apply our method to the Hybrid Mouse Diversity Panel (HMDP) collected in Bennett et al. (2010) study [6]. The Bennett et al. (2010) study consists of 25 phenotypes, 894 animals, and 98 strains. In this experiment, we impute body fat (BF) mass, which we consider as the target phenotype, by utilizing metabolic phenotypes (HDL, TG, TC, UC, FFA, and GLU) as the related phenotypes. The BF phenotype is measured by nuclear magnetic resonance (NMR). We assume

the BF phenotype is collected for only 200 animals, which is used as a training dataset to compute the pairwise correlations (see Figure S.6). The correlation between the imputed phenotype and the true BF phenotype is $r_{imp} = 0.4$. We perform similar experiments for this study to those performed on the TG phenotype for the NFBC dataset. Table 2 indicates the significant SNPs which pass our significant threshold of 0.05 for both imputed and real test datasets. This result indicates results similar to the NFBC dataset. For all of the variants the test statistic (Z-score) for the imputed phenotype is close to r_{imp} times the test statistic (Z-score) at the actual variant as shown in the last column of Table 2.

rsID	Real test data				Imputed test data				$\lvert Z_{imp}$ - 0.4 * $Z_{real}\rvert$
	β	se(β)	Z-score (Z_{real})	P-value	β	se(β)	Z-score (Z_{imp})	P-value	
rs38946050	-0.247	0.05887	-4.200	3.04E-05	-0.093	0.03220	-2.891	0.003	1.211
rs37558901	-0.163	0.03803	-4.286	2.09E-05	-0.051	0.0209	-2.448	0.01	0.733
rs27178379	-0.185	0.04433	-4.176	3.36E-05	-0.055	0.02435	-2.275	0.02	0.604
rs50810977	-0.163	0.03803	-4.286	2.09E-05	-0.051	0.0209	-2.448	0.01	0.733
rs51148868	-0.185	0.04433	-4.176	3.36E-05	-0.055	0.02435	-2.275	0.02	0.604
rs32339557	-0.163	0.03803	-4.286	2.09E-05	-0.051	0.02093	-2.448	0.01	0.733
rs51646366	-0.163	0.03803	-4.286	2.09E-05	-0.051	0.02093	-2.448	0.01	0.733
rs31560659	-0.163	0.03803	-4.286	2.09E-05	-0.051	0.02093	-2.448	0.01	0.733
rs50923350	-0.163	0.03803	-4.286	2.09E-05	-0.051	0.02093	-2.448	0.01	0.733
rs37193394	-0.205	0.04742	-4.331	1.72E-05	-0.056	0.02599	-2.161	0.03	0.428
rs26890141	-0.185	0.04433	-4.1769	3.36E-05	-0.055	0.02435	-2.275	0.02	0.604
rs46913800	-0.185	0.04433	-4.1769	3.36E-05	-0.055	0.02435	-2.275	0.02	0.604
rs38214662	-0.163	0.03803	-4.2867	2.09E-05	-0.051	0.02093	-2.448	0.01	0.733
rs47384543	-0.185	0.04433	-4.1769	3.36E-05	-0.055	0.02435	-2.275	0.02	0.604
rs51585751	-0.163	0.03803	-4.2867	2.09E-05	-0.051	0.02093	-2.448	0.01	0.733
rs29268223	-0.185	0.04433	-4.1769	3.36E-05	-0.055	0.02435	-2.275	0.02	0.604

Table 2: Comparison between the association test for BF phenotype on the real test data and the imputed test data in the HMDP.

3.5 Evaluating Imputation Power by Simulation

We evaluate the power of phenotype imputation through simulations. We remove the phenotype of interest from the dataset, then apply phenotype imputation to predict its value and measure the corresponding association power after imputation. In order to robustly measure this power, we randomize the individuals from whom we remove phenotype values.

Specifically, we follow the following simulation procedure. We consider a locus that has a significant association. In the first step, we compute the number of individuals that we need to remove their phenotypic values to obtain the statistical power of 50% for that locus. Let k indicate the number of individuals obtained from this step. In the second step, we randomly choose k

individuals and consider the phenotypic values for these k individuals that are missing. Then, we use our imputation model to impute the phenotypic values of these k individuals. We perform association test on the complete dataset. We repeat the second step 10,000 times in order to compute the statistical power. We compute the statistical power as the number of times where the computed association statistic is significant (with $p < 10^{-6}$). If the imputation is working, we would expect to see an increase in the power over 50%. We refer to the statistical power of 50% as the statistical power before imputation. We compute the value of k in the first step by randomly removing phenotypes of k individuals for 10,000 simulations. Then, for this value of k, we check whether the number of simulations where the association statistics is significant (with $p < 10^{-6}$) is 5,000 (50% of total simulations that corresponds to statistical power of 50%). We use TG, BMI, and SBP phenotypes from NFBC data to perform the power simulation. As shown in Table 3, the power gained by imputing the missing phenotype is $8\% - 33\%$.

Phenotype	rsID	Power after imputation	Power before imputation	Absolute power gain
TG	rs673548	83.59%	50%	33.59%
	rs10096633	62.16%	50%	12.16%
	rs3923037	63.74%	50%	13.74%
	rs6728178	80.97%	50%	30.97%
	rs6754295	76.40%	50%	26.40%
	rs676210	82.16%	50%	32.16%
BMI	rs987237	63.12%	50%	13.12%
	rs11759809	61.33%	50%	11.33%
SBP	rs782586	82.52%	50%	32.52%
	rs782588	81.72%	50%	31.72%
	rs782602	81.99%	50%	31.99%
	rs2627759	74.05%	50%	24.05%
	rs9791555	58.77%	50%	8.77%
	rs7799346	58.63%	50%	8.63%

Table 3: Measuring power of imputation by simulation in the NFBC data.

In the method section, we provide an optimal weight to combine imputed and observed summary statistics in a fixed effect meta-analysis. This process is beneficial when we have access to the summary statistics. We utilize the same simulation process described above. We randomly pick k individuals to mask them as individuals with missing phenotypes. Then, we compute the summary statistics (s_c) for individuals whose phenotypic values are observed. We impute the missing phenotypes and compute the summary statistics ($\hat{s}_m$) for individuals whose phenotypic values

are missing. To combine these statistics, we have two options. First option, we use Equation (17) to combine the computed summary statistics in an optimal way. We refer to this option as imputation-based fixed-effect meta-analysis. Second option, we use fixed effect meta-analysis with typical fixed-effect meta-analysis weights. In this case, we use $w_c = \sqrt{n_c}$ and $w_m = \sqrt{n_m}$. We refer to this option as general fixed-effect meta-analysis. We observe that, using the second option in which the weights are not optimal, we have loss of power (see Table 4). Moreover, we compare the first option, which is optimal, to the previous simulations where we combine the imputed and observed phenotypic values. Finally, we compute the summary statistics. We observe there is a small difference between these two cases. In this experiment we use TG phenotype from NFBC dataset.

rsID	Imputation-based Fixed-effect Meta-analysis Power	General fixed-effect meta-analysis Power
rs673548	83.56%	82.30%
rs10096633	62.14%	45%
rs3923037	63.65%	60.86%
rs6728178	80.96%	80.00%
rs6754295	75.49%	74.31%
rs676210	82.01%	80.85%

Table 4: The optimal meta-analysis strategy to combine summary statistics for imputed and observed phenotype achieves maximum power. Imputation-based Fixed-effect Meta-analysis uses the optimal weights that is shown in Equation (17). General fixed-effect meta-analysis uses the typical fixed-effect meta-analysis weights where the weight for each study is square root of the number of samples in that study.

The statistical power of imputation depends on r_{imp} which is the correlation between the imputed and true phenotype (see Figure 3). We use similar experiments as described above. We consider imputing TG phenotype using HDL, LDL, CRP, and GLU phenotypes. There are $2^4 - 1 = 15$ possible combinations for these four phenotypes to impute the TG phenotype (excluding one combination that refers to a case where none of the four phenotypes are used for imputation). For each combination of phenotypes, we compute r_{imp} and the statistical power for a given variant. In Figure 3, the black circle indicates one of the 15 possible combinations for imputing TG phenotype. The x-axis is the computed r_{imp} for a given combination of phenotypes, and the y-axis is the computed statistical power. The red curve indicates a second order polynomial that is fitted to the black circles. We observe that the statistical power increases as we increase the value of r_{imp} (see Figure 3). There are two factors that increase r_{imp}. First factor to increase r_{imp} is the number

of phenotypes that satisfies the NMM assumption. As we use more phenotypes that satisfy the NMM assumption in our imputation method, we can increase r_{imp} that result in increases of power. Second factor to increase r_{imp} is the correlation between phenotypes that are used to impute target phenotype. As we use more correlated phenotypes, we can increase r_{imp} that result in increases of power.

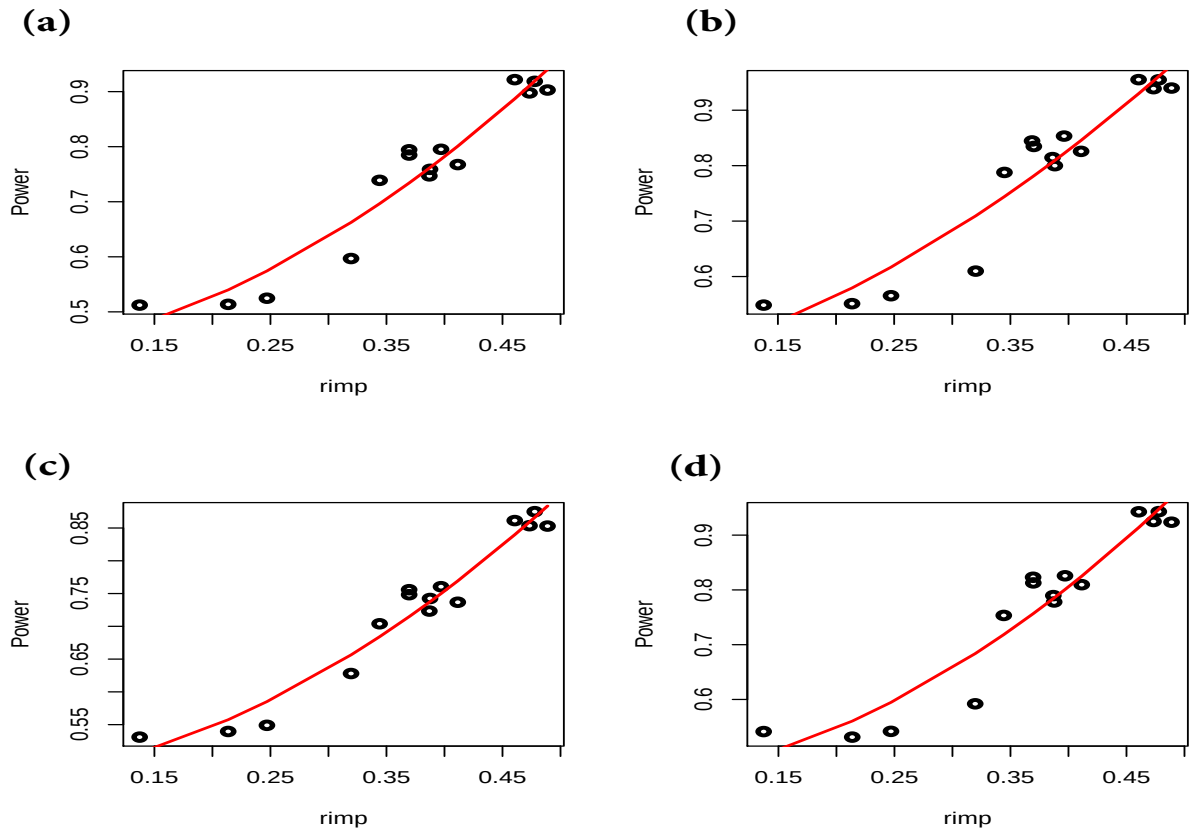

Figure 3: Increase of r_{imp} increases the statistical power. The x-axis is the r_{imp} and the y-axis is computed power. Panels (a), (b), (c), and (d) illustrate the effect of r_{imp} on power of imputing the TG phenotype for rs6728178, rs673548, rs6754295, and rs676210, respectively. We impute the TG phenotype in NFBC data using HDL, LDL, CRP, and GLU phenotypes. The black circle indicates the r_{imp} and the statistical power for a combination of four phenotypes to impute TG for one variant. The red curve indicates a second order polynomial which is fitted to the black circles.

3.6 Utilizing Simulation Data to Validate Our Model

In the method section, we show the r_{imp}, which is the correlation between imputed and true phenotype, is equal to $\sqrt{R_{\neg\ell\ell}^T \Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}}$. We impute one of the phenotypes by utilizing any combination of the remaining nine phenotypes. There are $2^9 - 1$ possible combinations for these nine phenotypes

to impute the desired phenotype in NFBC dataset. We compute the difference between r_{imp} and $\sqrt{R_{\neg\ell\ell}^T \Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}}$. We observe that this difference is small (see Figure S.5). We compute r_{imp} as a correlation between imputed and true phenotype. We perform this experiment for all the nine phenotypes (TG, HDL, LDL, BMI, CRP, GLU, INS, SBP, and DBP) in NFBC dataset.

Next, we compare the difference between the computed association statistics for imputed phenotype and the analytical association statistics obtained from Equation (14). We simulated phenotypes for 1,000, 5,000, and 10,000 individuals and we considered three, four, five, and six phenotypes in each simulation. We simulate multi-phenotypes utilizing the matrix-variate that is used in previous works [14, 52–54]. We run each of the simulations for 10,000 times and our result is the average of 10,000 runs. The result of these simulations are shown in Table S.3.

4 Discussion

We propose a novel method for the problem of phenotype imputation. The primary advantage of our framework is it increases power of GWAS on phenotypes that are difficult to collect. We provide an analytical power computation that allows researchers to prospectively determine the benefit of the imputation for a given dataset. Another advantage of our method is that it allows the use of summary statistics when the raw genotypes are not available.

Our model assumes that the phenotypes follow a normal distribution. This assumption is widely accepted in the GWAS community [20, 24, 54]. When the phenotypes are not normal, one possible way is to transform the phenotypes to follow a normal. In our case, we applied the inverse normal transform to the data which is heavily used by many studies [3, 36, 43] and verified that when all of the phenotypes in the NFBC data were transformed, the phenotypes as a set followed a multivariate normal distribution (see Figure S.2). Another possible way to deal with non-normal phenotypes is to use the weighted combination of statistics approach. Asymptotically, the multivariate central limit theorem applies if the datasets are large enough and the statistics themselves will follow a multivariate normal distribution. Thus, using weighted combination of z-scores will still control the type I error, although its optimal properties may not be guaranteed for non-normal phenotypes.

Our framework is closely related to the noisy measurement model (NMM) in that both the power calculation and the connection to weighted combination of statistics are based on NMM. In

Methods, we showed that we can assume a more complex polygenic model, and NMM is equivalent to polygenic model where we assume that the genetic correlation is the same as environmental correlation. For situations that this is not the case, we also developed weighted combination of statistics approach that is optimized for polygenic model. If we have an accurate estimate of genetic and environmental correlations, using this approach may show better performance. However, estimating genetic correlations using SNP data often requires thousands of individuals. By contrast, the phenotypic correlations can be accurately measured relatively easily from a much smaller set of individuals. Therefore, we expect that our standard solution based on phenotypic correlation and NMM will be a practical solution for situations that the size of complete dataset is small. Moreover, our analysis based on real data shows that NMM is a reasonable model for most of loci that we evaluated.

An implicit assumption of our approach is that we expect that we can borrow information of target phenotype from the proxy phenotypes. That is, we assume that there will be pleiotropy between phenotypes that are reflected in correlations. If this is not the case, such as the TG-associated locus (rs1260326) that were not at all associated to other phenotypes, the power to detect such a locus using other phenotypes is considerably limited. Note that this is not the limitation of only our method, but can be a limitation of any possible approaches that depend on proxy phenotypes. Nevertheless, our NFBC analysis shows that such situation is relative rare (one out of seven loci) compared to the situations that our method was effective.

It is worth mentioning that phenotype imputation has some similarities to phenotype prediction. In phenotype prediction, one typically predicts phenotypes based on available genetic information. One of the widely used methods for phenotype prediction is BLUP (Best Linear Unbiased Prediction) [19]. Phenotype prediction is an active research area, and various approaches have been proposed to solve this problem efficiently [32, 34]. The main difference between phenotype prediction and phenotype imputation lies in the main goal of the approaches. The main goal of phenotype prediction is to have a method that predicts the phenotypic values as close as possible to the true value using the genetic data and possibly using other phenotypes. However, in phenotype imputation, the goal is to impute the phenotypic values using other phenotypes such that we can recover the associated signals had we have collected the imputed phenotype. Therefore, we can not use the genetic data for phenotype imputation. If we utilize the genetic data in our imputation, we would

not be able to perform genetic association because the genetic data would be used twice (once in imputation and once again in the GWAS).

Phenotype imputation is in several ways analogous to genotype imputation [8, 22, 23, 26, 30]. In genotype imputation, we want to impute the missing genotypes. As in phenotype imputation, if we use one tagged variant in the genotype imputation to impute the missing variant, we lack sufficient power when we perform GWAS on the imputed genotype. However, if we use a panel of reference individuals and multiple variants, we can achieve higher power. This is similar to our phenotype imputation where utilization of multiple phenotypes will achieve higher power than only using one phenotype. These similarities are the reasons we use the name "phenotype imputation" for this problem.

Our method controls type I errors even in the scenario that there are systematic differences between the reference (first dataset) and target (second dataset) datasets. In the case of systematic differences, power will be affected, but our method will not report false positives.

We acknowledge the fact that more sophisticated machine learning can be utilized, including techniques such as support vector machines (SVM) [9], LASSO [47], Elastic-net [55], and supervised PCA [4] to solve the phenotype imputation problem and improve the imputation power. Moreover, these methods do not make any assumption on the distribution of collected phenotypes. However, these methods are designed for general missing data problems and do not utilize the genetic data. Recently, a multiple imputation method [10] is proposed that incorporate the genetic similarity (kinship) between individuals to perform phenotype imputation. This method performs better than generalized machine learning methods described above. However, all of these methods require access to individuals' raw data which in most cases is not possible. This is one the main advantages of our method that we can perform imputation using available summary statistics. In addition, we provide an analytical power calculation for our method, while performing analytical power computation is not easy for other methods.

In our approach, due to our parametric assumptions, we know the exact distribution of the imputed phenotype and can directly use the mean value of this distribution as the imputed value. Furthermore, we utilize the variance of the missing phenotype in our analysis of the statistical power. If we are using a more sophisticated machine learning method for the imputation, such as the methods mentioned above, we can use multiple imputation techniques [39, 40] to obtain the

confidence intervals for the imputation.

5 Acknowledgements

F.H., E.Y.K, M.B. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589 and 1331176, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. B.H. is supported by the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea (2015-0222) and the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (HI14C1731). S.M. and C.V. are supported by the National Institutes of Health grant R01-GM083198-01A1. E. E. is supported in part by the NIH BD2K award, U54EB020403. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691). The authors declare that they have no competing interests.

References

[1] Abecasis, G. R., Noguchi, E., Heinzmann, A., *et al.* (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *The American Journal of Human Genetics*, **68**(1), 191–197.

[2] Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, **55**(1), 193–196.

[3] Ardlie, K. G., Deluca, D. S., Segrè, A. V., *et al.* (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235), 648–660.

[4] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**(473), 119–137.

[5] Balise, R. R., Chen, Y., Dite, G., *et al.* (2007). Imputation of missing ages in pedigree data. *Human Heredity*, **63**(3-4), 168–174.

[6] Bennett, B. J., Farber, C. R., Orozco, L., *et al.* (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome research*, **20**(2), 281–290.

[7] Bobb, J. F., Scharfstein, D. O., Daniels, M. J., Collins, F. S., and Kelada, S. (2011). Multiple imputation of missing phenotype data for qtl mapping. *Statistical Applications in Genetics and Molecular Biology*, **10**(1), Article 29.

[8] Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics*, **124**(5), 439–450.

[9] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297.

[10] Dahl, A., Iotchkova, V., Baud, A., *et al.* (2016). A multiple-phenotype imputation method for genetic studies. *Nature Genetics*.

[11] de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., *et al.* (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, **17**(R2), R122–8.

[12] Dunning, A. M., Durocher, F., Healey, C. S., *et al.* (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *The American Journal of Human Genetics*, **67**(6), 1544–1554.

[13] Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome research*, **18**(4), 653–660.

[14] Furlotte, N. A. and Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics*, **200**(1), 59–68.

[15] Ghosh, S. (1988). Statistical analysis with missing data. *Technometrics*, **30**(4), 455–455.

[16] Gordon, T., Castelli, W. P., Hjortland, M. C., Kannel, W. B., and Dawber, T. R. (1977). High density lipoprotein as a protective factor against coronary heart disease: the framingham study. *The American journal of medicine*, **62**(5), 707–714.

[17] Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, **5**(4), e1000456.

[18] Han, B., Hackel, B. M., and Eskin, E. (2010). Postassociation cleaning using linkage disequilibrium information. *Genetic Epidemiology*, **35**(1), 1–10.

[19] Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, **1973**(Symposium), 10–41.

[20] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**(2), 497–508.

[21] Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. *Bioinformatics*, **31**(12), i206–i213.

[22] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, **44**(8), 955–959.

[23] Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**(6), e1000529.

[24] Kang, H. M., Sul, J. H., Service, S. K., *et al.* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**(4), 348–54.

[25] Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**(2), 139–144.

[26] Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**(8), 816–834.

[27] Lippert, C., Listgarte, J., Liu, Y., *et al.* (2011). Fast linear mixed models for genome-wide association studies. *Nature Methods*, **8**(7), 833835.

[28] Listgarten, J., Lippert, C., Kadie, C. M., *et al.* (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, **9**(7), 525526.

[29] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley-Blackwell.

[30] Marchini, J. and Howie, B. (2008). Comparing algorithms for genotype imputation. *The American Journal of Human Genetics*, **83**(4), 535–9; author reply 539–40.

[31] McCulloch, C., Searle, S., and Neuhaus, J. (2011). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. Wiley.

[32] Meuwissen, T. and Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, **185**(2), 623–631.

[33] Meyer-Lindenberg, A. and Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, **7**(10), 818–827.

[34] Ober, U., Ayroles, J. F., Stone, E. A., *et al.* (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in drosophila melanogaster. *PLoS Genetics*, **8**(5), e1002685.

[35] Ohashi, J. and Tokunaga, K. (2001). The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using snp markers. *Journal of Human Genetics*, **46**(8), 478–482.

[36] Okada, Y., Kubo, M., Ohmiya, H., *et al.* (2012). Common variants at cdkal1 and klf9 are associated with body mass index in east asian populations. *Nature Genetics*, **44**(3), 302–306.

[37] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

[38] Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, **69**(1), 1–14.

[39] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

[40] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

[41] Sabatti, C., Service, S. K., Hartikainen, A.-L., *et al.* (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, **41**(1), 35–46.

[42] Schunkert, H., Knig, I. R., Kathiresan, S., *et al.* (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, **43**(4), 333–338.

[43] Speliotes, E. K., Yerges-Armstrong, L. M., Wu, J., *et al.* (2011). Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genetics*, **7**(3), e1001324.

[44] Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, **5**(5), e1000477.

[45] Sterne, J. A., White, I. R., Carlin, J. B., *et al.* (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **338**.

[46] Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**(2), 367–383.

[47] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[48] Vaitsiakhovich, T., Drichel, D., Angisch, M., *et al.* (2014). Analysis of the progression of systolic blood pressure using imputation of missing phenotype values. In *BMC Proceedings*, volume 8, page S83. BioMed Central Ltd.

[49] Voight, B. F., Scott, L. J., Steinthorsdottir, V., *et al.* (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, **42**(7), 579–89.

[50] Willer, C. J., Speliotes, E. K., Loos, R. J. F., *et al.* (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, **41**(1), 25–34.

[51] Zaitlen, N. and Eskin, E. (2010). Imputation aware meta-analysis of genome-wide association studies. *Genetic Epidemiology*, **34**(6), 537–542.

[52] Zhou, J. J., Cho, M. H., Lange, C., *et al.* (2015). Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Human Heredity*, **79**(2), 93–104.

[53] Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, **11**(4), 407–409.

[54] Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, **9**(2), e1003264.

[55] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

Appendix

Phenotype	r_{imp}
TG	0.58
HDL	0.41
LDL	0.34
BMI	0.61
CRP	0.37
GLU	0.33
INS	0.62
SBP	0.63
DBP	0.63
Height	0.10

Table S.1: The r_{imp} computed for the 10 phenotypes using the other nine phenotypes in NFBC dataset.

Phenotype imputation for cases where different subsets of phenotypes are missing

In the method section, we explained our method, as the target phenotype is the only missing phenotype. Unfortunately, when the number of related phenotypes is large there exist many individuals where one or more phenotypic values are missing. Let $\boldsymbol{c}$ indicate a vector of size $\ell-1$ where each element of the vector is zero or one value. Vector $\boldsymbol{c}$ indicates which phenotypes are missing excluding the target phenotype, the i-th element of $\boldsymbol{c}$ is one for the cases where the i-th phenotype is missing. We refer to $\boldsymbol{c}$ as one configuration of missing phenotypes in the second dataset. Given we have $\ell-1$ phenotypes, we have at most $2^{\ell-1}$ such configurations. Let $\mathcal{C}$ indicate the set of all possible configurations, $\mathcal{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots\boldsymbol{c}_{2^{\ell-1}}\}$. Let $Y^{(2)}_{\boldsymbol{c_i}}$ indicate a new partition of the second dataset to set

Phenotype	β	se(β)	Z-score	P-value
TG	-0.037	0.006	-5.37	**8.159e-08**
HDL	0.005	0.005	0.984	0.325
LDL	-0.012	0.013	-0.977	0.328
BMI	-0.035	0.022	-1.594	0.110
CRP	0.001	0.001	1.1737	0.240
GLU	0.003	0.005	0.699	0.484
SBP	0.028	0.189	0.152	0.879
DBP	0.008	0.166	0.050	0.959
Height	0.032	0.095	0.344	0.730

Table S.2: rs1260326 violates the NMM assumption. Association results for rs1260326 on the ten phenotypes in the NFBC data obtained from EMMAX [24]. As shown, this SNP is significant for TG phenotype. However, the p-value of this SNP for other phenotypes is extremely high.

of individuals which miss exactly the phenotypes denoted by configuration $\boldsymbol{c}_i$. We can easily extend our method to impute the target phenotype for these individuals that belong to configuration $\boldsymbol{c}_i$ by removing the phenotypes that are missing for these individuals. Thus, $\Sigma_{\neg\ell}$ and $R_{\neg\ell\ell}$ are computed similarly as mentioned in previous section while we excludes the phenotypes which are missing for these individuals. Then, we apply Equation (7) and Equation (14) to compute the imputed target phenotype and the imputed marginal statistics respectively for only these individuals utilizing the observed phenotypes. It is possible, we can have up to $2^{\ell-1}$ different configurations, thus, we can have up to $2^{\ell-1}$ different marginal statistics for each configuration. Let $\hat{s}_{\boldsymbol{c}_i}$ indicate the imputed marginal statistics for the configuration $\boldsymbol{c}_i$. Then, we compute the total marginal statistics by applying the fixed-effect meta-analysis as shown in previous section. Thus, we have:

$$\hat{s}_\ell = \frac{w_1\hat{s}_{\boldsymbol{c}_1} + w_2\hat{s}_{\boldsymbol{c}_2} + \cdots + w_{2^{\ell-1}}\hat{s}_{\boldsymbol{c}_{2^{\ell-1}}}}{\sqrt{w_1^2 + w_2^2 \cdots w_{2^{\ell-1}}^2}} \tag{27}$$

where w_i is the optimal weight for the marginal statistics for the configuration $\boldsymbol{c}_i$ which is proportional to correlation between the imputed target phenotypic values and the true uncollected phenotypic values for all the individuals in configuration $\boldsymbol{c}_i$.

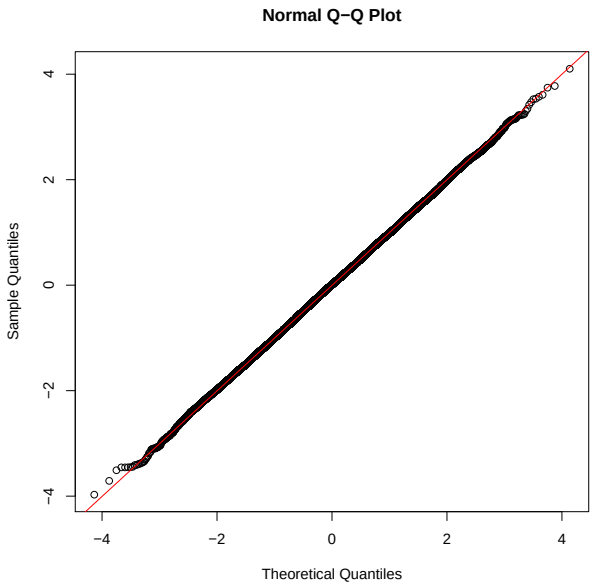

Figure S.1: Q-Qplot of imputed TG phenotype from NFBC dataset. The x-axis is the theoretical distribution which is standard normal distribution and y-axis is the computed z-score.

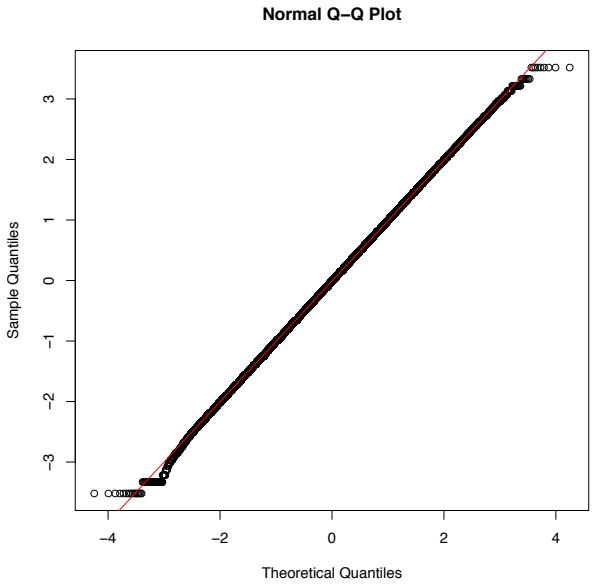

Figure S.2: Q-Qplot of the transformed phenotypes from the NFBC dataset. The x-axis is the theoretical distribution which is standard normal distribution and y-axis is the transformed phenotypes from the NFBC dataset. We used the inverse normal transformation on all the 10 phenotypes. We merge all the 10 phenotypes to one large phenotype and plot the Q-Qplot. This indicates that transformed phenotypes as a set follows a multivariate normal.

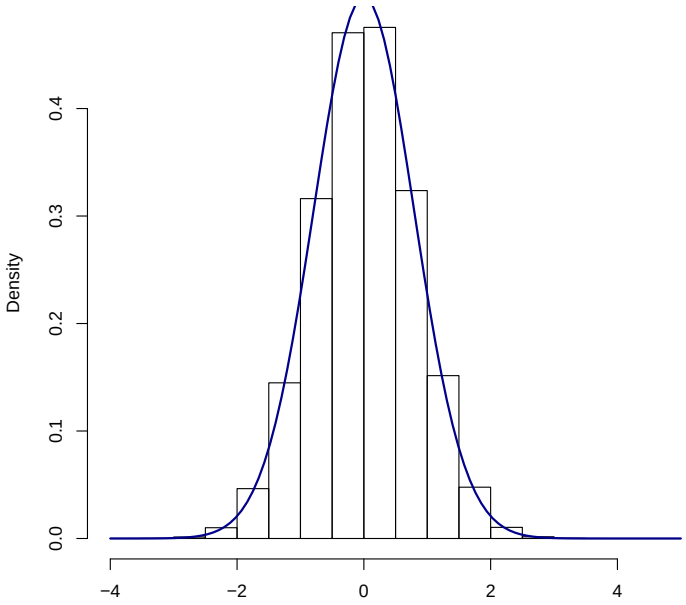

Figure S.3: Difference between the imputed marginal statistics and analytical marginal statistics for BMI phenotype. Imputed marginal statistics is obtained from the association between the genotype and the imputed phenotype and analytical marginal statistics is equal to the marginal statistics computed on the true target phenotype scaled by r_{imp}. The blue curve is the normal distribution with mean zero and variance 1-r^2_{imp}. This histogram indicates this difference follows a normal distribution (mean zero and variance 1-r^2_{imp}). Thus, for most null variants the NMM assumption holds.

#individuals	Number of phenotypes			
	3-phen	4-phen	5-phen	6-phen
1000	0.100 ± 0.079	0.087 ± 0.071	0.086 ± 0.068	0.081 ± 0.064
5000	0.043 ± 0.033	0.040 ± 0.031	0.037 ± 0.028	0.034 ± 0.026
10000	0.032 ± 0.024	0.028 ± 0.022	0.023 ± 0.017	0.022 ± 0.015

Table S.3: The residual difference between the computed association statistics for the imputed phenotype and the analytical association statistics under the case the NMM assumption holds.

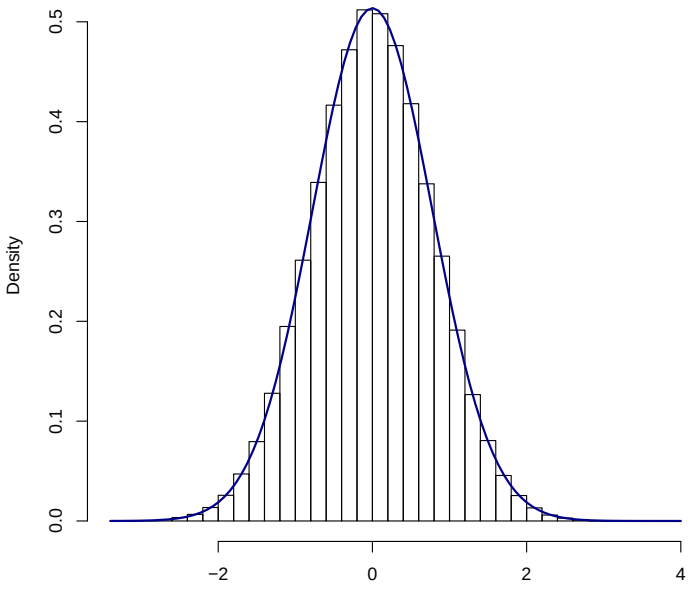

Figure S.4: Difference between the imputed marginal statistics and analytical marginal statistics for SBP phenotype. Imputed marginal statistics is obtained from the association between the genotype and the imputed phenotype and analytical marginal statistics is equal to the marginal statistics computed on the true target phenotype scaled by r_{imp}. The blue curve is the normal distribution with mean zero and variance 1-r^2_{imp}. This histogram indicates this difference follows a normal distribution (mean zero and variance 1-r^2_{imp}). Thus, for most null variants the NMM assumption holds.

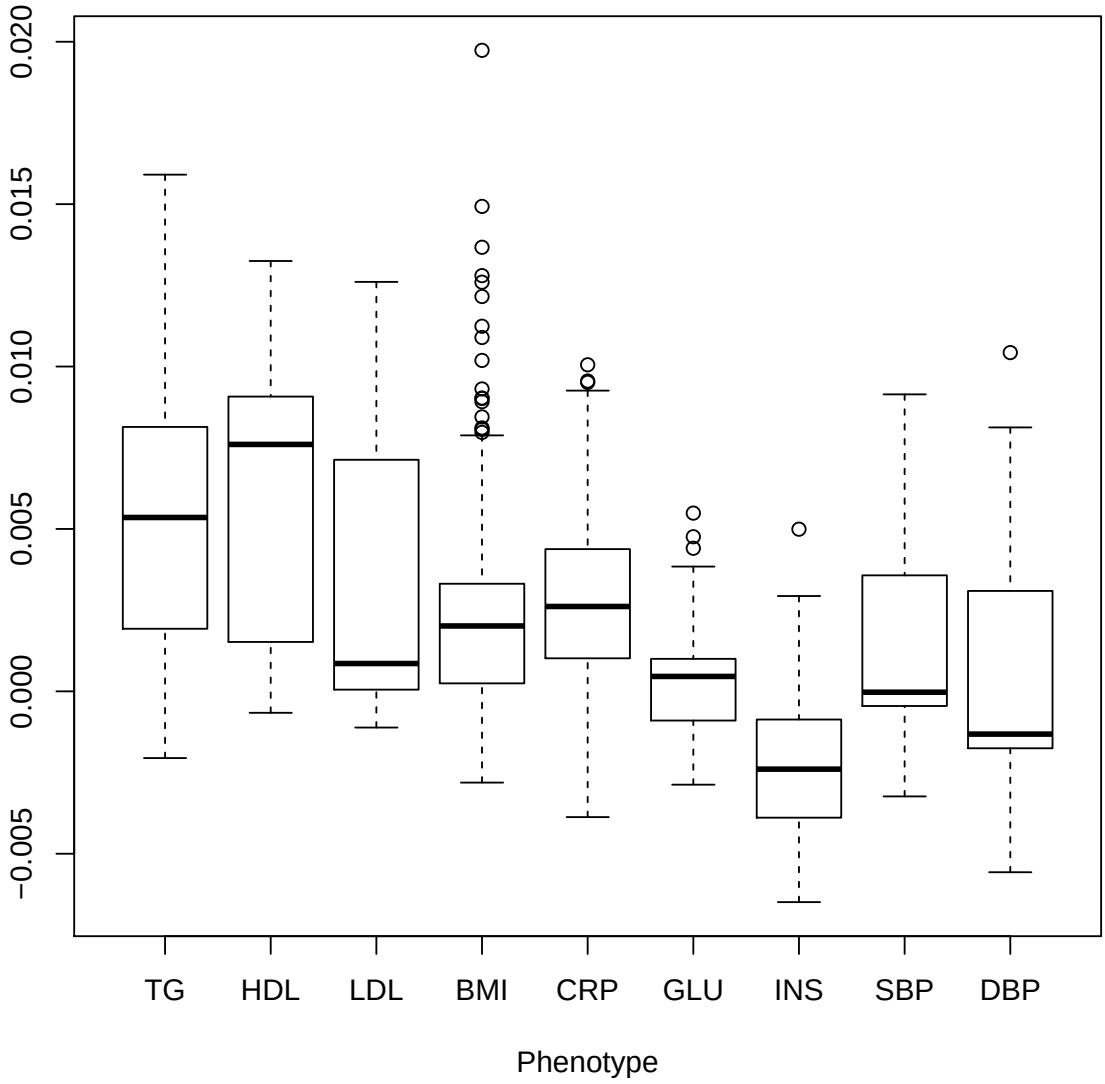

Figure S.5: Difference between analytical and empirical value of r_{imp} is small. The x-axis is the phenotype which we impute using the remaining nine phenotypes from NFBC dataset. The y-axis is the difference between empirical and analytical value of r_{imp}. The empirical value is the correlation that is computed between true and imputed phenotype. The analytical value of r_{imp} is $\sqrt{R_{\neg\ell\ell}^T \Sigma_{\neg\ell}^{-1} R_{\neg\ell\ell}}$.

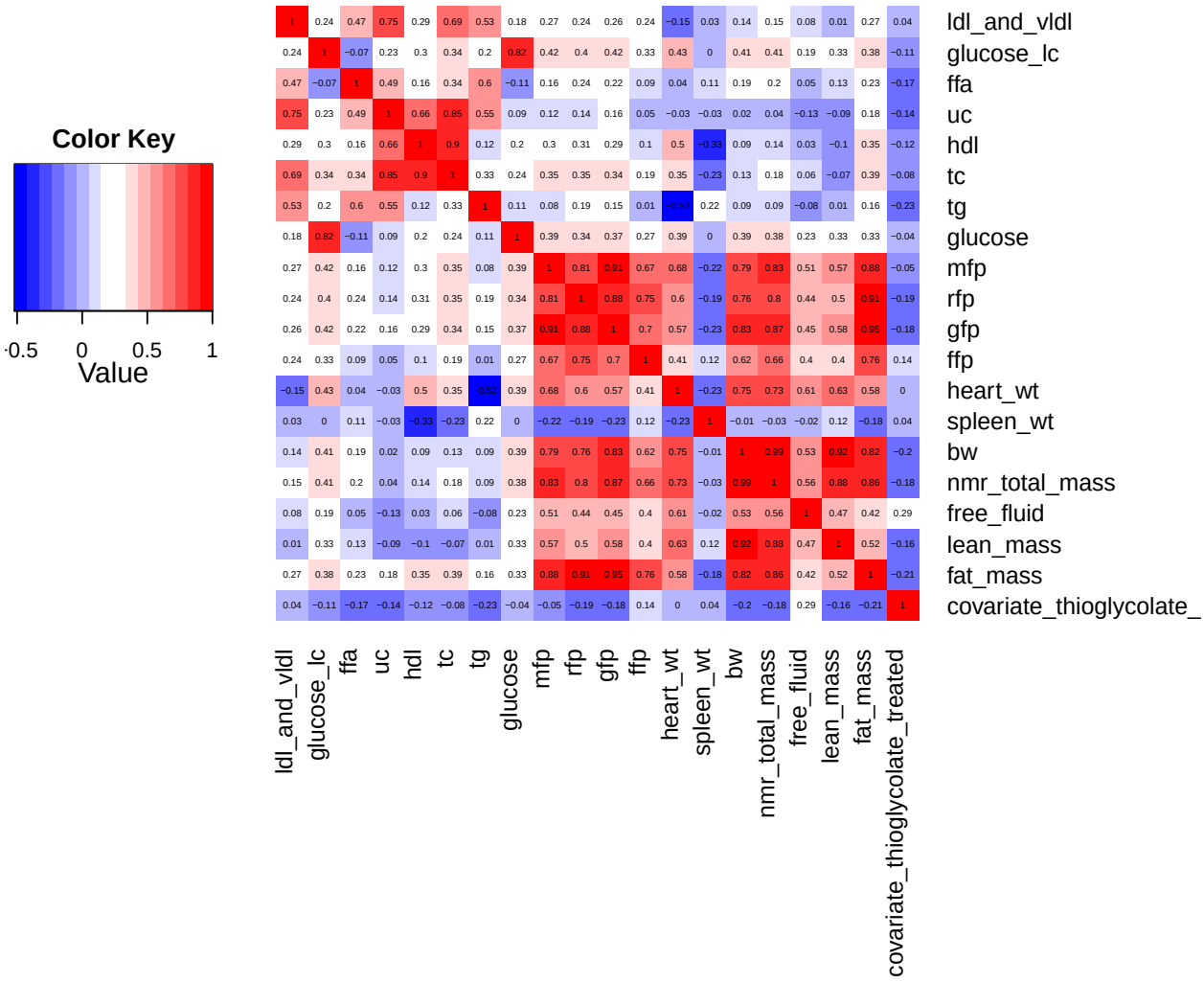

Figure S.6: The pairwise correlation between each pair of phenotype in the HMDP dataset.